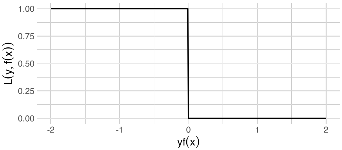
Introduction to Machine Learning

Advanced Risk Minimization
Classification and 0-1-Loss



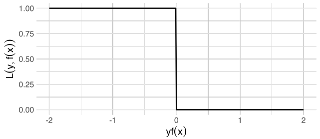
Learning goals

- 0-1 loss
- Risk minimizer(s) / Optim. predictions
- Bayes error rate
- Generative approach in classification

0-1-LOSS

- Discrete classifier $h : \mathcal{X} \to \mathcal{Y}$
- Maybe most "natural": 0-1-loss

$$L(y, h(\mathbf{x})) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}}$$



- For $\mathcal{Y} \in \{-1, +1\}$ and scoring classifier $f(\mathbf{x})$ can write it in terms of margin $\nu = yf(\mathbf{x})$

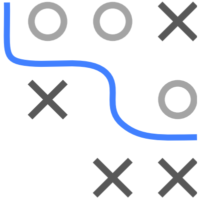
$$L(y, f(\mathbf{x})) = \mathbb{1}_{\{\nu < 0\}} = \mathbb{1}_{\{yf(\mathbf{x}) < 0\}}$$

- For $\mathcal{Y} \in \{0, 1\}$ and prob. classifier $\pi(\mathbf{x})$

$$L(y, \pi(\mathbf{x})) = y\mathbb{1}_{\{\pi(\mathbf{x}) < 0.5\}} + (1 - y)\mathbb{1}_{\{\pi(\mathbf{x}) \geq 0.5\}} = \mathbb{1}_{\{(2y-1)(\pi(\mathbf{x}) - 0.5) < 0\}}$$

- Analytical properties: Not continuous, even for linear $f(\mathbf{x})$ optim. problem NP-hard = close to intractable ▸ Feldman et al. 2012

RISK MINIMIZER FOR DISCRETE CLASSIFIERS



- Again, unravel with law of total expectation (works for multiclass)

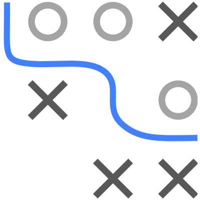
$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}_{xy}\left[L\left(y, f(\mathbf{x})\right)\right] = \mathbb{E}_x\left[\mathbb{E}_{y|x}[L\left(y, f(\mathbf{x})\right)]\right] \\ &= \mathbb{E}_x\left[\sum_{k \in \mathcal{Y}} L(k, f(\mathbf{x}))\mathbb{P}(y = k \mid \mathbf{x})\right]\end{aligned}$$

- $\eta_k(x) := \mathbb{P}(y = k|\mathbf{x})$ is true posterior probability for class k
- For binary case, we denote $\eta(\mathbf{x}) := \mathbb{P}(y = 1 \mid \mathbf{x})$ and get:

$$\mathcal{R}(f) = \mathbb{E}_x\left[L(1, \pi(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(0, \pi(\mathbf{x})) \cdot (1 - \eta(\mathbf{x}))\right]$$

- Above formulas work for any loss, not only 0-1; and hard labelers, scorers or prob. classifiers
- Especially for hard labelers and arbitrary misclassif. costs, we see: produces cost-optimal decision, weighted by posterior probs; (we see this again in cost-senslearning)

0-1-LOSS: OPTIMAL PREDICTIONS

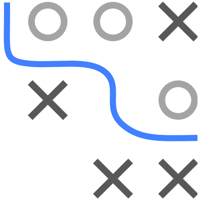


- For multiclass and hard labeler $h(\mathbf{x})$
- Optimal constant

$$
\begin{aligned}
h_c^* &= \arg\min_{k \in \mathcal{Y}} \sum_{l \in \mathcal{Y}} L(l, k) \cdot \mathbb{P}(y = l) \\
&= \arg\min_{k \in \mathcal{Y}} \sum_{k \neq l} \mathbb{P}(y = l) \\
&= \arg\min_{k \in \mathcal{Y}} 1 - \mathbb{P}(y = k) \\
&= \arg\max_{k \in \mathcal{Y}} \mathbb{P}(y = k)
\end{aligned}
$$

- Translation: Predict most probable class
- Empirical version: $\hat{h}_c = \arg\max_{k \in \mathcal{Y}} \hat{\pi}_k$
- Risk minimizer / optim. cond. prediction / Bayes optim. classifier:
  $h^*(\tilde{\mathbf{x}}) = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(y = k \mid \mathbf{x} = \tilde{\mathbf{x}})$

BAYES RISK / BAYES ERROR RATE



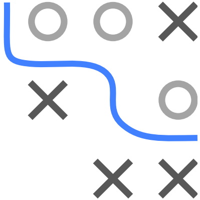
$$\mathcal{R}^* = 1 - \mathbb{E}_x \left[\max_{k \in \mathcal{Y}} \mathbb{P}(y = k \mid \mathbf{x}) \right]$$

For binary case, can write risk minimizer and Bayes risk as:

$$h^*(\mathbf{x}) = \begin{cases} 1 & \eta(\mathbf{x}) \geq \frac{1}{2} \\ 0 & \eta(\mathbf{x}) < \frac{1}{2} \end{cases}$$

$$\mathcal{R}^* = \mathbb{E}_x \left[\min(\eta(\mathbf{x}), 1 - \eta(\mathbf{x})) \right] = 1 - \mathbb{E}_x \left[\max(\eta(\mathbf{x}), 1 - \eta(\mathbf{x})) \right]$$

GENERATIVE CLASSIFIERS



- So, $\arg\max_{k \in \mathcal{Y}} \mathbb{P}(y = k \mid \mathbf{x} = \tilde{\mathbf{x}})$ is what we want to do
- Assume we can model densities given classes and use Bayes:

$$\mathbb{P}(y = k \mid \mathbf{x} = \tilde{\mathbf{x}}) = \frac{p(\tilde{\mathbf{x}}|y = k)\mathbb{P}(y = k)}{p(\tilde{\mathbf{x}})}$$

- Then

$$\arg\max_{k \in \mathcal{Y}} \mathbb{P}(y = k \mid \mathbf{x} = \tilde{\mathbf{x}}) = \arg\max_{k \in \mathcal{Y}} p(\tilde{\mathbf{x}}|y = k)\mathbb{P}(y = k)$$

- Then we can estimate these conditional densities and the prior probs, and classify via them
- This idea we will see in so-called “generative approaches” for classification, so in LDA, QDA, etc.

EXAMPLE

- Assume $\mathbb{P}(y = 1) = \frac{1}{2}$
- And conditional densities of x per class as normal
$$\begin{cases} \phi_{\mu_1,\sigma^2}(x) & \text{for } y = 0 \\ \phi_{\mu_2,\sigma^2}(x) & \text{for } y = 1 \end{cases}$$
- Bayes optimal classifier = orange; Bayes error = red